MASTER'S THESIS

Data Governance:

A conceptual framework in order to prevent your Data Lake from becoming a Data Swamp

Charikleia Paschalidi
2015

Master (120 credits)
Master of Science in Information Security

Luleå University of Technology
Department of Computer science, Electrical and Space engineering

Data Governance:
A conceptual framework in order to prevent your Data Lake from becoming a Data Swamp

Paschalidi Charikleia

Master Program
Master of Science in Information Security

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering

COPYRIGHT

Contact Information:
Project Member:
Paschalidi Charikleia
E-mail: chapas-2@student.ltu.se

University Advisor:
Devinder Thapa
E-Mail: devinder.thapa@ltu.se

LULEÅ UNIVERSITY OF TECHNOLOGY
Department of Computer & Electrical Engineering
Division of Computer and System Science
SE-971 87 Luleå
SWEDEN

CONTENTS

Abstract:

Information Security nowadays is becoming a very popular subject of discussion among both academics and organizations. Proper Data Governance is the first step to an effective Information Security policy. As a consequence, more and more organizations are now switching their approach to data, considering them as assets, in order to get as much value as possible out of it. Living in an IT-driven world makes a lot of researchers to approach Data Governance by borrowing IT Governance frameworks.

The aim of this thesis is to contribute to this research by doing an Action Research in a big Financial Institution in the Netherlands that is currently releasing a Data Lake where all the data will be gathered and stored in a secure way. During this research a framework on implementing a proper Data Governance into the Data Lake is introduced.

The results were promising and indicate that under specific circumstances, this framework could be very beneficial not only for this specific institution, but for every organisation that would like to avoid confusions and apply Data Governance into their tasks.

Keywords:

Data Governance, Data Lake, Standards, Policies, Business, Risk and Compliance, IBM tooling, Hadoop. Data Integration Layer (DIL), Data Quality, Scrum, Agile Methods.

1 Introduction

Data is often described as the oil of the 21st century. According to IBM, "approximately more than three exabytes of digital data are created daily around the world" (Zhu et al., IBM, 2011). In the age of smartphone and social media, the exposure to data is so huge that we are drowning in it. Therefore, companies understood how significant big data is and, they are switching their focus from software and hardware to the data they process, in order to gain competitive advantage in knowing more than their competitors (Kemp IT Law, 2015, p. 1-169).

Big Data refers to the use of the huge volumes of data, its collection and analysis generated by our digital lives. A few years ago, companies either did not have access or they did not know what to do with it. Nowadays, though, companies use it to reveal new insights into the Business Processes, Demand or Customer Sentiment, as well as to anticipate and react to changes. Banks, for instance, are using Big Data in order to get a better insight of their customers, so that they can be able to provide them with better services. (Taylor, Nevitt & Carnie, 2012).

However, having access to data is not enough. According to Dave Coplin, Director of Research and Futurologist at Microsoft, "Data itself is worthless. It is what you do with it that has value." In order to explore the value of data, we need to explore the data first. This, requires the recognition of data as corporate assets and the implementation of some form of Data Governance for an effective Data Management.

The purpose of this thesis is to introduce a road map about how to develop a framework for the Data Governance of Big Data. Since most of the knowledge that exists nowadays comes from IT Governance, this thesis aims to separate those two fields and focus on Data Governance itself.

1.1Problem Area

The problem is derived from the increased importance for businesses to focus more and more on their data. It is usually common, for the majority of the organisations, although they begin to realize the importance of their data that they do not know what to do with it. Nowadays, when our digital life is as intense and important as our normal everyday life, and the exposure of mountains of data becomes bigger, companies need to know what their "Big Data" is, as well as how to use it efficiently.

According to Financial Times, "Companies find data useful, because they can reveal new insights into Business Processes, Demand or Customer Sentiment that help them anticipate or react to changes". Banks, for instance, use Big Data in order to obtain a more understandable picture concerning their customers. In that way, they can deliver more appropriate and customized services to their clients, as well as spot fraud easier (Taylor, Nevitt & Carnie, 2012). Therefore, the need for Data Governance is currently more important than ever.

In this paper, there is an attempt to construct a framework for implementing Data Governance in a Big Financial Institution in the Netherlands, since they are trying to implement a new Data Lake solution for Commercial Banking. Currently, there is no such framework implemented and there is a lot of confusion inside the company about who is responsible for the data and what they should do about it.

1.2Objectives and current situation

The objective of this paper is to come up with a conceptual framework that will help a specific Dutch Financial Institution, successfully implement Data Governance on their newly existing Data Lake for Commercial Banking.

The company is a leading Commercial Bank in Belgium, Netherlands and Luxemburg as well as in Central and Eastern Europe and has strong global franchises in Specialized Finance and Financial Markets. Its clients are supported through an extensive network in more than 40 countries and its headquarters are in Amsterdam, the Netherlands. The Bank's customer base includes individuals, small and medium-sized businesses, large corporations, institutions and governments.

Currently, the Commercial Banking (CB) is releasing the GDIL (Global Data Integration Layer) which is the first version of the CB Data Lake, focused on IT-Business interaction, with the help and tools of IBM Company. DIL is a central post-transaction data integration hub for Commercial Banking, rather than a point-to-point solution for every application. It is supposed to reduce the interface complexity and customization needs on packages and it is both real-time and batch-based. In addition, DIL should not be considered as a wide data warehouse, since it doesn't keep historical data.

Ideally, the data in a bank are well-organized, but in reality, it is challenging to find the right data and get it. What happens so far within the institution is, a creation of copies of data needed in order to send it to the departments who request it. The main goal of the Data Lake is to get rid of those copies that are both expensive and time-consuming.

Building the Data Lake will be demand-driven, based on user cases that pop up. In other words, every project that arises and requires Data Lake functionality, will contribute to the implementation of the Data Lake solution.

The foundation components consist of:

- Data Glossary and Data Catalogue
- Enterprise Data Overlay tooling
- Security tooling
- Spikes on Real-time streaming and analysis, an NoSQL

The whole Data Lake solution consists of three parts, each of them with its own responsibilities:

1 **System of Records**, responsible for:
  - The quality of the data
  - Delivering Daily Data feed to the Data Lake

- Defining the transformation to a common Esperanto language
- Archiving of data

2 **Data Lake** itself, responsible for:

- Receiving data from the System of Records
- Delivering data to the Receiving System
- Implementing the transformations to/from the common Esperanto language

3 **Receiving System**, responsible for:

- Data consolidation
- Receiving (daily) data feed from the Data Lake
- Defining the transformation from the common Esperanto language
- Building historic data

1.3Motivation to research on this topic

The reason I would like to do a research on this subject is that it is a very interesting and well promising topic which already plays a significant role in the business world. Proposing a conceptual framework will give a starting point not only to the Financial Institution, but to any other company that does not know how to start with implementing Data Governance as well. This is because, although organizations start focusing more and more on their data and I believe that in a few years, the need for a company to protect its data will become even more crucial, there is still a gap between understanding the need of Data Governance and knowing where to start from.

What is more, as an employee of a big financial Institution in the Netherlands and as a member of the DIL team, I was assigned to do a research on the topic and present it to the rest of the team, as well as to come up with a conceptual framework that will help the team to properly implement Data Governance into the Data Lake.

1.4 Research Question

Based on the objectives mentioned above, the main research question of this thesis is the one below:

- *How should a conceptual framework be built in order to prevent a Data Lake from becoming a Data Swamp?*

In order to manage to answer the research question, I firstly need to do a research in data governance, in order to find out what it is already known on the topic in both the academic and business world.

The second step will be to explore what it is already known inside the company on the subject, find out the gaps and fill them in.

Later on, I intend to research the tools IBM provides in order to successfully implement the Data Governance into the Data Lake.

The next step, which is also the practical part of the thesis, is to interview both sides, those who send as well as those who receive data from the data lake in order to understand what data are meaningful and should exist in the lake.

Eventually, taking into consideration the steps above, I will need to come up with a conceptual framework and define roles in every part of the lake that will help to successfully implement Data Governance to the Data Lake.

1.5Assumptions and Delimitations

1.5.1 Assumptions

The study assumes that people who are currently working on either side of the Data Lake in the company are willing to cooperate and answer the questions asked honestly and within the time-frame of the assignment. Another assumption is that the steps can be aligned with the Scum methodology in order to be properly implemented into the everyday activities of the team.

1.5.2 Delimitations

As mentioned above, there is only limited research on Data Governance. Most of researches either transfer knowledge form IT Governance to Data Governance or use resources from practitioners, analysts and consultants, due to lack of academic resources. As a result, collecting all the data needed and coming up with a specific questionnaire might be challenging.

In addition, people working in the team should be willing to adjust to the new framework and the agile way of working in order to manage to implement the steps suggested in the theoretical framework.

2 Literature Review

2.1 Literature Review Method

The literature review of this paper is based on the already existing literature about Scrum Methodology, Data repositories, Data Management and Data Governance. Since there is not a lot of literature on the topic, as it was mentioned above, resources from practitioners, analysts and consultants will be also used in order to come to meaningful conclusions.

During the process of the research, an analysis of the questionnaires will take place, in order to understand people's awareness about Data Governance in the organisation, as well as the level of process of Data Governance the organisation itself is in. Later on, I will be actively involved in the process of implementing Data Governance in the Data Lake as a member of the DIL department and I will come up and suggest a conceptual framework for implementing Data Governance.

The following chapter of this research contains the description of the methodology that was used during the development of the thesis. Here, the approach taken by the author is defined, as well as the way the approach was executed.

2.1.1 Purpose of the literature review

The purpose of the literature review is:

- To demonstrate the value of Data Governance in the Organizations
- To identify previous Data Governance Frameworks
- To discuss the problems or limitations that exist when it comes on implementing a proper Data Governance Framework to a specific organisation in order to find a way to overcome them.

2.1.2 Searching for the literature

The research for this thesis started by using more generic terms like "Data Governance", "Data management", "Data Lake", "Big Data", etc. The tools used for our research was Google Scholar, the University of Lulea Library PRIMO database and "ResearchGate", a social networking site for researchers to share papers, ask and answer questions, and find collaborations. The search for literature was stopped when the material found started presenting repetitive patterns (Leedy, P. D., &Ormrod, J. E., 2005)

2.1.3 Practical screen

The quantity and variety of the results required the focus on results only applicable to the specific situation the DIL team is currently facing. The need for Data Governance in the newly formed department was crucial and the lack of literature on the topic made the situation even more challenging. There is not a lot of literature concerning Data Governance yet, although the need for it is getting bigger and bigger over time. Most of the knowledge about data Governance is borrowed from the "IT Governance" field and therefore, it is not accurate or complete. (ACM Journal of Data and Information Quality, One Size does not fit all, Vol. 1, No. 1, Article 4, Pub. date: June 2009)

Another aspect that had to be taken into consideration is that the DIL, and the whole Financial Institution is working according to the Scrum Agile method and, as a consequence, every task they are performing should be also aligned with Scrum Methodology. So far, there is no Data Governance Framework aligned with the Scrum Agile method in literature.

After getting enough knowledge from a variety of articles and papers concerning Governance and Big Data and after experiencing myself the way of working in the team, I had to come up with a Data Governance Framework applicable to the specific department and align the procedure to the Scrum methodology. I therefore, decided to disregard articles that would not add value to this thesis due to their specific content. Therefore, the papers selected for this research were chosen my reviewing the titles and the abstracts.

2.2 Traditional Data Governance Practices

The reason for using Data Governance is to ensure the quality, integrity and security in an organization. There are several methods to develop Data Governance and it is usually up to each organization to decide the exact way of applying it. The most Traditional Data Governance Practices are the following (http://agiledata.org/essays/dataGovernance.html#Traditional) :

1. **Valued Corporate Assets**. What is important for organizations is to make it as easy as possible for the DevOps teams to comply to the corporate IT infrastructure as well as to take advantage of it. Therefore, guidance, metadata definitions, and reusable assets such as frameworks and components, are being adopted in this method in case they are perceived to add value to developers. When data standards are sensible, understandable, and easily accessible, there is a significantly greater chance that people will actually follow them, while forcing them to do so, will have the opposite results.
2. **Scenario-Driven Development**. This method uses scenarios in order for developers get the whole picture of the system by understanding how people are going to use it and therefore, try to meet the people's requirements..
3. **Include data professionals as active participants on development teams**. This will enable the people working on a project to develop a "team mentality" and cooperate to reach the maximum result possible. This is one of the fundamental concepts of the Agile Data method.
4. **Educate developers**. Developers need to understand the importants and benefits of the Data Management effort in order to try their best for achieving a good result. Once they know why something needs to be done, and how to do it effectively, their performance could be increased.
5. **Adapt the Process**. It is a fact that teams vary in size, distribution, purpose, criticality, need for oversight, and skills. This means that not everything can be done the same way by every person or team and therefore, the approach to support data-oriented activities, including governance, will vary by team.
6. **Align Team Structure With Architecture**. The organization of the data team should always reflect the architectural structure of the organization.

7. **Align HR Policies With IT Values**. There must be specific rewards appropriate for the mindset of your technical staff in order to ensure that they follow your data governance strategy.
8. **Align Stakeholder Policies and IT Values**. Since the development efforts are driven by the stakeholders, the stakeholders from their side must be realistic in their demands on IT, and understand the implications of their decisions.
9. **Business-Driven Project Pipeline**. There is a need for investment in the IT activities that are well-aligned to the business direction, and match with the priorities of the enterprise.
10. **Embedded Compliance**. Compliance should be built into the day-to-day processes instead of having a separate compliance process which often results in unnecessary overhead.
11. **Flexible Architectures**. The architectures that are component-based, service-oriented, or object-oriented and implement common architectural and design patterns lend themselves to greater levels of consistency, reuse, and adaptability.
12. **Pragmatic Governance Body**. Effective governance bodies focus on enabling development teams in a way that is both cost and time effective.
13. **Promote Self-Organizing Teams**. The most suitable people for planning the work that must be done, are the ones who are actually going to do it. This doesn’t mean that the team should be out of control, but it definitely has to plan its own tasks.
14. **Risk-Based Milestones**. Since mitigation of the risks of your project is crucial having throughout your project several milestones that address to those risks that teams work toward is important.

2.3 Problems – Limitations of current methods

The Traditional Data Governance Methods mentioned above, often suffer from certain common problems – limitations. Those limitations are mentioned below:

1 Data governance doesn't usually fit into the overall IT governance effort.

2 Data governance efforts are ignored. Development teams often prefer to “work around” an organization’s data group. In order for data governance efforts to succeed, development teams should truly collaborate.

3 Data governance is too difficult to conceive. Usually, development teams report that the data group within their organization is too difficult to work with.

4 Data governors are often too slow to respond. As a consequence, developers tend to believe that what they believe is best.

5 Data governors **are not considered to provide** value**. This is** because of the additional bureaucracy involved with traditional approaches.

3 Research Method

The method used to answer the question is mostly qualitative. Qualitative research involves the use of qualitative data, such as interviews, documents, and participant observation data to understand and explain social phenomena.

According to Myers (2009), qualitative research methods were developed in the social sciences to enable researchers to study social and cultural phenomena (Myers, M. D., June 1997, pp. 241-242). Examples of qualitative methods are action research, case study research and ethnography. Qualitative data sources include observation and participant observation (fieldwork), interviews and questionnaires, documents and texts, and the researcher's impressions and reactions.

The reason I decided to do qualitative research is that this was the most appropriate method for the research. This is because; qualitative methods are designed to help researchers understand people and the sociocultural areas within which they live. According to Kaplan and Maxwell (1994), the goal of understanding a phenomenon from the point of view of the participants and its particular social and institutional context is largely lost when textual data are quantified (Kaplan, B. and Maxwell, J.A., 1994, p.45-68)

There are several different qualitative research methods which require different skills, assumptions and practices. Some of the different research methods are the action research, the case study research, the ethnography and the grounded theory.

The first one, Action Research, according to Rapoport (1970), "aims to contribute both to the practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework" (Rapoport, R.N., 1970, pp. 499-513). More specifically, Avison et al (1999) describe this method as: "an iterative process involving researchers and practitioners acting together on a particular cycle of activities, including problem diagnosis, action intervention, and reflective learning". (p. 94) Case Study Research is defined by Yin (2002) as an empirical study that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident (Yin, R. K., 2002).

There are many different Qualitative Research methods that were initially taken into consideration in order to perform this study. At first, a Case Study Research seemed the most appropriate one. The main reason for that was the fact that, at the beginning of the project, a lot of interviews took place in order to get an understanding of people's comprehension of Data Governance and participation in the Data Lake.

Later on, though, during the study process, Case Study was proved to be an unsuitable method, mainly because of the nature of the research. The goal of the study was not only observation but also, intervening and active participation in the creation of the Framework needed in order to implement Data Governance into the DIL department of the Financial Institution.

Therefore, the most suitable method for this particular study was considered to be the *Action Research*. The main reasons that led me to this conclusion are the following:

1. My main goal was to come up with results that would be beneficial not only for me and my research but also for the company itself, since it refers to an existing and interesting case. The research is being executed within the Financial Institution, in a real-life context and in collaboration with practitioners.

2. The knowledge that I would expect to gain from the study would be immediately applied into the organisation. The main purpose of the research was to actively participate in the creation of a framework that would help in an effective Data Governance implementation in the Data Lake.

3. Therefore, in order to achieve that, I would have to work close with the members of the department as part of the team and make observations of people´s conception and understanding when it comes to Data Governance. At the same time, I would get immediate feedback from stakeholders in order to improve the framework in progress.

4. In most cases, conventional information systems development methodologies and not very effective when it comes to real life cases. This is mainly because of the complexity of the organisation, as the researcher cannot easily understand and develop an effective solution just by observing, sitting on his/her desk. Action Research though, helps the researcher to test the concepts in real life cases, gain feedback from the team and align with it accordingly. Tis way he/she can refine and improve the framework in process.

5. Last but not least, consensus from the Financial Institution side was ensured in order to perform the thesis, as well as easy access to the people who were interviewed, since I am currently working in this organisation and I am part of the team.

3.1 Action Research Methodology Phases

According to Susman and Evered (1987), Action Rresearch consists of 5 phases that follow an iterative process. Those phases are the ones described below:

1. Diagnosing: In this initial phase, the researcher needs to understand the company’s current situation and be familiarized with the internal way of working. This is the most crucial phase of the Action Research since this is the starting point based on which the proposed changes and framework was created. The diagnostic phase is described in more details in chapter 5.1 and 5.2, explaining the current situation of the Financial Institute as well as the reason there is a need for the Data Lake.

2. Action planning: The action planning phase involves the actions necessary to solve the problems the Institution had at the beginning of the research. Those problems were identified in the Diagnosing Phase, guided by the theoretical framework (Chapter 4). In other words planning establishes the target for change and the approach to change (Baskerville, 1999, p.15). The actions needed to implement Data Governance into the Data Lake are described in section 5.3, where the IBM launch approach and tooling are introduced as a way of implementing Data Governance in a proper way.

3. Action taking: In this phase, the actual implementation of the proposed framework proved to be a challenge due to confusions and misunderstandings in between the team members. Therefore, this phase consists mostly of interviews, trainings, meetings and further research in order to understand people's confusion and lack of understanding, gather information about what actions should be taken and exchange of opinions about the right approach of the suggested framework.

In addition, the team adopted the Agile Scrum Methodology of working, something that increased transparency of the team since everybody had to daily inform the rest of the team members of his or her tasks. As a consequence, although this transparency caused a lot of confusion at the beginning, it helped and increased the teams collaboration over time and helped the team use the right people for the right task.

4. Evaluating: The part of the interview aimed to evaluate the people's view, inside and outside the team, about Data Governance. A clear understanding of what people who work for DIL, as well as parties who come in contact with the Data Lake, think of the solution provided and presented in section 5.4.

In addition, an effort to create a profile of each participant's prior knowledge of Big Data, Data Governance, Data management, as well as their personal involvement with the Data Lake, in order to be able to interpret their views on the proposed framework more effectively.

5. Specifying learning: In this phase, a reflection of the previous steps needs to be made, as the actual results of the processed need to be finalized. The results of the research seemed quite promising and a framework on both a higher and a lower level was introduced in chapter 6.

3.2Action Research Challenges

Usually, Action Research in not the safest choice to perform a study since the researcher has to deal with many challenges described below:

a) It is a time consuming, demanding and risky method as it involves the production of empirical data and the project might not evolve as initially planned

b) It personally demanding, as it requires a constant engagement and commitment between the researcher and the practitioners

c) It might not be the appropriate solution in case the researcher has only limited amount of time on his disposal. (Simonsen, 2009, p. 10)

4 Theoretical Framework

4.1Big Data

There are, so far, various definitions for Big Data, since the phenomenon is both of technical and sociological nature. In most cases, people use the term to describe the mountains of data they have been exposed to, mostly generated by our digital lives. It often involves the collection and analysis of data that have high volume and low value information. The main three characteristics of Big Data are the so-called 3 Vs; volume, velocity and diversity (McAfee, A., and Brynjolfsson, E., 2012, p.59-66), although a fourth V is introduced by the Oxford Internet Institute; Veracity (Oxford Internet Institute, 2014). Those data seem insignificant if we look at each piece of data separately, but its meaning increased dramatically when we look at them as a whole.

Although no standard definition about Big Data exist, nobody can deny its existence. According to Nicola Askhan, Data Governance Coach, people think of data the same way they think of the air. They take it for granted and they do not understand its added value until they lose it.

Nowadays, Big Data is everywhere; in every online login, in every use of application, in any online purchase, etc. (CRISAN, ZBUCHEA& MORARU, 2014). According to Boyd and Crawford, "We live in the age of Big Data" (Boyd, D. and K. Crawford., 2012, p. 662-279) Those data are of massive scale and complexity.

The advantages of Big Data are, among others, the following (Oxford Internet Institute, 2014):

- Advocating and Facilitating: The fact that Big Data can be massively produced in real time and without any restrictions in size, provides the advantage of showing the granular detail of a given problem and therefore, help in creating interactive tools and engaging the people involved in the problem so that they get a deep understanding of the situation.

- Describing and predicting: Nowadays, researchers can combine social data and real-time information that have never been combined before in order to provide high-resolution, dynamic data sources and methods of analysis.

- Facilitating Information Exchange: The use of particular social networks used to connect people with organizations provides companies with the ability to learn more about their clients.

- Accountability and Transparency: Big Data, if manager correctly, help us get a better insight on the information we need and facilitate accountability and transparency in real life.

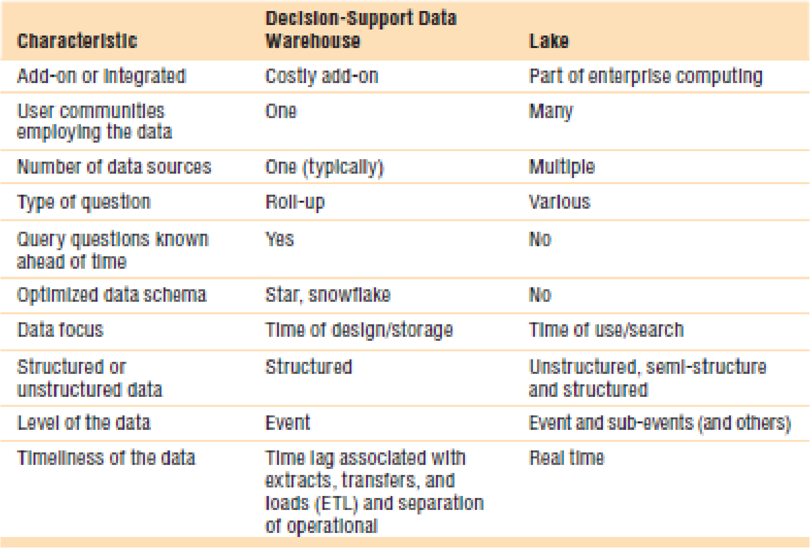
As a consequence, there are some concerns that usually arise when it comes to Big Data. Those concerns have to do mainly with the quality of data analysis and the protection of privacy and intimacy. Therefore, the governance of Data is of great importance.

4.2 Data Lake – What is it?

As mentioned above, Big Data is everywhere and companies have just started to understand its added value. There are, therefore, a lot of concerns about protecting it, since organisations need to keep it safe and of a good quality in order to add value to their business. The Data Lake is a way to keep all the data safe as well as maintain a single point of truth.

More specifically, Data Lake is a new concept, used to describe the evolutionary form of data repositories. It is a thinking framework as well as a technical solution and it defines how to receive and deliver data, how to separate them into different repositories based on their special characteristics, and how to secure them. In addition, Data Lake also defines how to access data for either reporting or exploration and analytic modelling, and the tools and the way to implement them.

The concept of the data lake is that data will be delivered once, avoiding the recreation of copies across departments and divisions within a company. A combination a real-time and batch delivery of data exists, a support for unstructured data is provided when using new technologies, and new technologies enable bringing the analytics to the data instead of bringing the data to the analytics. The main differences between the traditional Data Warehouses and the Data Lake are described in table 1 (Daniel E. O'Leary, 2014).



Characteristic	Decision-Support Data Warehouse	Lake
Add-on or integrated	Costly add-on	Part of enterprise computing
User communities employing the data	One	Many
Number of data sources	One (typically)	Multiple
Type of question	Roll-up	Various
Query questions known ahead of time	Yes	No
Optimized data schema	Star, snowflake	No
Data focus	Time of design/storage	Time of use/search
Structured or unstructured data	Structured	Unstructured, semi-structure and structured
Level of the data	Event	Event and sub-events (and others)
Timeliness of the data	Time lag associated with extracts, transfers, and loads (ETL) and separation of operational	Real time

Table 1: Data Warehouse versus Data Lake (Daniel E. O'Leary, 2014)

As illustrated in figure 3, the Data Lake works as follows:

- Receives data from the System of Records, as well as from other Data Lakes in their native format.
- Implements the transformation in an Esperanto language, according to IBM Data Base Workbench, which has been already decided to be the English language.
- Delivers data to the Receiving System in their preferred/native format.
- The new sources describe other information besides the data managed by the system of records (internal sources, log files from customer interactions, etc.)
- Systems used by data scientists and business analysts and their rules and models are described in the Decision Model Management.
- Information sources that need to be shared are handled by the Information Owner.
- There is also a Governance, Risk and Compliance platform used to demonstrate compliance on regulations and business policies.
- Line of Business Insight applications are designed to provide reports, search and simple analytics capabilities that are being controlled by the business.
- In addition, there is a Data Lake Operations team responsible for managing the data lake operations.

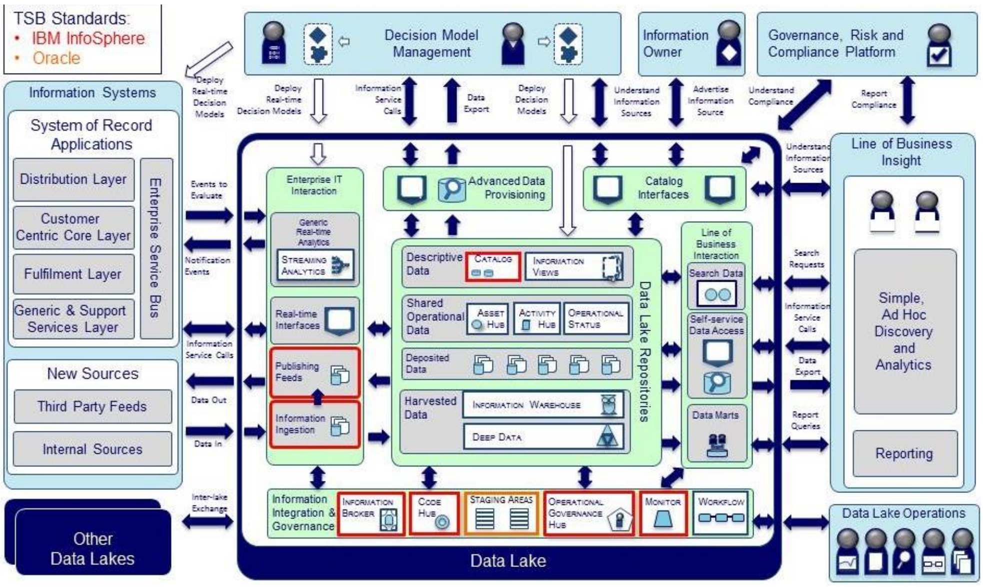


Figure 3: Data Lake Architecture, IBM Corporation

As we can see from the picture above, the Data Lake includes the following capabilities:

- Generic Real-time Analytics: This feature enables complex processing of events as well as real-time analytics based on the data that is coming into the lake.
- Enterprise-IT interaction: The interaction between business and IT and, the synchronization of data in and out of the Data Lake is taking place in this component.
- Information Integration & Governance: The control of mechanisms for managing, governing and transforming all information in the Data Lake is being provided here.
- Line of Business Interaction: This feature offers easy access to data in the lake for business users.
- Harvested Data: It contains repositories where data is integrated, combined and enriched in order to enable Decision Model Management.
- Deposited Data: A warehouse that allows users to store their data temporarily to facilitate them. This data might later be added in the Data Lake or be removed. These warehouses are governed by the lake.
- Shared Operational Data Store: A set of warehouses that provide the operational state of the financial institution as well as a historical view of all data and activities from the Systems of Records that feed the Data Lake. It also allows real-time access to data in the Data Lake.
- Descriptive Data: A warehouse that contains metadata (= data about data) of relevant data assets that may be in or outside the Data Lake.
- Catalog Interfaces: It provides metadata information, including details of the information collection, meaning and type of information, and the profile of the information values within each information collection.
- Advanced Data Provisioning: It refers to the capability of the system to provide ad-hoc sandbox environment that includes data for advanced analytics purposes.

4.3 Data Management

The official definition provided by DAMA International, the professional organization for those in the data management profession, is: "Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise." In other words, Data Management is the process of making and implementing decisions concerning data. More specifically, it involves determining the actual metrics employed for data quality (Wende, 2007), by focusing on collecting, organizing, storing, processing and presenting high-quality data. Since it is addressing technical, as well as organizational issues, there is a need for Data Management to integrate business and IT functions (Weber, Otto, & Osterle, 2009).

In the case of the Financial Institution, Data Management is a key building block to support its strategy. It is the objective to ensure that the company has trustworthy data that can enable:

- **Next Generation Digital Bank:** Data is an asset, which is equally important as people and systems in order to be the next generation digital bank.

- **Enhanced Customer Experience:** Offer customized services to the customers, make them feel that their data is safe, etc

- **Accurate Reporting** (both internal and external)

- **Operational Excellence:** Lower operational and IT costs

In our case, except Data Management, we also have to concentrate on Metadata Management. Metadata is information concerning our data and they are crucial in order to understand the value of our initial, raw data. IBM Corporation refers to Metadata Management as a procedure of cataloguing information about data objects. Since the majority of the organizations do not know how to make use of their data, they tend to spread them across different systems and departments. As a consequence, users who own, access and manage those systems cannot communicate with each other easily. The definition given by IBM Redbooks is that "Metadata management refers to the tools, processes, and environment that are provided so that organizations can reliably and easily share, locate, and retrieve information from these systems." ("Jackie Zhu et al., Metadata Management with IBM InfoSphere Information Server, IBM Redbooks, October 2011)

In other words, metadata management is the necessary process that organizations need in order to make sure that the final reports and analysis are coming from the right data sources, complete and with high quality. It is the tools, processes and environment that are provided to enable an organisation to answer the Question, "How do we know what we know about our data?" ("Jackie Zhu et al., Metadata Management with IBM InfoSphere Information Server, IBM Redbooks, October 2011"). In order to achieve that, metadata management has to be provided in every step of the implementation process.

4.4 Data Governance

Data Governance, is part of Data Management and addresses the issues above within corporate structures. It is a set of procedures that makes sure that crucial information is properly managed throughout the organization, encompassing professionals from both IT and business departments. In addition, it ensures that data itself is trustworthy and in case of poor data quality, only people are to blame. What is more, Data Governance reflects the enterprise's stage of evolution when it comes to data and the power and intelligence it adds to the company. (Sarsfield, S., 2015)

There are two dominant theories about Data Governance. The first one, the so called "Modernist Theory", supports that within any policy domain, the structure of the institution is the one that shapes the consideration of policies, the interaction of the actors, as well as the evaluation of the outcomes. The second one, the "imperative" approach, claims that governance theory itself should provide an explanation of the reasons on which people act as well as make sense of them. (Dr Alberto Asquer, 2013)

The aim of this thesis is to contribute on the “interpretive” approach to the theory of governance by investigating the understanding of Data Governance and the tooling provided within a particular organizational institution and coming up with a framework of data governance implementation. The main reason to contribute to the second scenario is that a theory of governance that heavily relies on institutions, structures, and processes seems well equipped to account for the reproduction or adjustment of social practices within a relatively stable context, where individual beliefs, preferences, and meanings may not significantly affect the established institutional and social order.

Usually, Data Governance is being handled by a group of professionals inside the company who form the Data Governance Team. The members of that team might differ in each organisation (Sarsfield, S., 2015). In case of the Financial Institution here is a newly formed team, called Global Data Management and suggests the following teams and roles for the upcoming solutions:

- **Global Data Council**: A team that has the mandate to agree upon or amend the data strategy framework, which comprises of the data governance, definitions, architecture, quality, processes, sustained awareness, lineage, organisation structure, privacy, security, etc. It consists of a mix of data suppliers, data users, data custodians, architect and CDO and meets every 3 months to discuss the solutions.

- **Operational Data Council**: Refers to the team that makes decisions concerning data management and steers programs on a tactical level. It consists of a mix of data suppliers, data users, data custodians and architects and they agree to meet once every two weeks.

- **Data owner:** The data owner is the one who is responsible and accountable (has legal rights and complete control over) for the data within his systems. He/she decides how the data is acquired and used. He/she is aware of where the data is further distributed within ING.

- **Data steward:** The Data steward is the one who assists the data owner in his day-to-day activities in order to ensure the data is managed and of the required quality as defined and agreed in ING’s data policies. The role of data steward is optional. It is at the discretion of the data owner whether or not to assign this role.

- **Data user:** The data user is the one who needs data from one or more data owners in order to aggregate, further process, store, or to create meaningful reports for different purposes (management reporting, regulatory reporting, statutory reporting, customer intelligence, etc.)

- **Data custodian:** Data custodian is the one who helps the data user by providing the organisation and systems to gather data from one or many data owners, store and further process as per the requirements of the data user.

- **Data definition owner:** The data definition owner is the one who defines standard definitions for the data attributes in the data categories he/she is responsible for, in order to facilitate their exchange across the entire organisation. He/she does so by carefully aligning the requirements of the data users and the data owners.

- **Data quality officer:** Data quality officer helps the data definition owner to set data quality requirements for every data element based on the requirements of the data users and translates them to the data owners and data stewards.

- **Data architect:** The data architect is the one who designs the data models and data lineage effectively and efficiently to ensure that the data policies and procedures are implemented and adhered to. He/she also identifies that redundant systems/processes that can be eliminated.

- **Data security officer:** Data security officer is the one who ensures that the data is secure in terms of storage, exchanges, etc. and that cybercrime threats are adequately taken care of.

- **Data privacy officer:**Data privacy officer is the one who ensures that the data exchange is always subjected to the privacy issues with respect to legal and other requirements between global and local exchange or local to local exchange.

Data Governance focuses on the following three areas:

- **Life-cycle management**: Management of the creation, storage, retention, recovery, and destruction of information that might be needed by the company and regulatory authorities.

- **Protection**: Control of data use in order to provide privacy and security.

- **Trusted source of information**: Assurance of the information quality, common understanding, timeliness, accuracy, and completeness of the information. (Zhu et al., 2011)

The types of technologies that are usually used in Data Governance are the following ((Sarsfield, S., 2015):

- **Preventative:** This technology stops bad quality data that shouldn't exist in the lake from coming into the organization. This way, any possible disruption is limited. Tools of this technology are type-ahead, workforce management and, data quality dashboard.

- **Diagnostic and health:** Organizations use this technology in cases when the damage is already done. The data exist in storage for many years already and therefore, there is need for data profiling as well as batch data quality.

- **Infrastructure:** The tools that could be used here are metadata, ETL, Master Data Management, enterprise-class data quality tools and data monitoring.

- **Enrichment:** The tools used here are services and data sources.

4.4.1 Data Governance Challenges

According to the findings by the IBM Data Governance findings, the most important challenges data governance faces nowadays are the following (The IBM Data Governance blueprint, Leveraging best practices and proven technologies, May 2007):

- Data Governance is usually inconsistent and leads to disconnection between IT and Business departments
- The Governance policies are not linked to structured requirements gathering and reporting
- The risks should be observed from a lifecycle perspective along with common data repositories, policies and standards
- Business glossaries and metadata should be used in order to fill the gap in semantic differences in global enterprises
- There are not enough technologies to assess data asset values that link security, privacy and compliance
- The controls and architecture used are being deployed before modeling the long-term consequences

In order for a company to be able to identify any Data Governance issues, The IBM Data Governance Council suggests six key questions that the organisation should answer (The IBM Data Governance blueprint, Leveraging best practices and proven technologies, May 2007):

(1) *Do you have a Governor?*
(2) *Have you surveyed your situation?*
(3) *Do you have a data governance strategy?*
(4) *Have you calculated the value of your data?*
(5) *Do you know the probability of the risk?*
(6) *Are you monitoring the efficacy of your controls?*

4.4.2 Data Governance first steps

No Data Governance program can start without having first someone accountable and with the authority to make things happen. This person, the governor, should create a Governance council and investigate the existing situation in order to create a vision of the level of Data Governance in the future. The next step would be to evaluate the data with the proper tools so that the organisation can enhance, protect or measure the data’s contribution to the business. Furthermore, a proper probability of risk estimation should be done since it is crucial for any enterprise to be aware of how its data might be abused and how often. Last but not least, a company should be constantly monitoring the efficacy of its controls since the organisation itself, its data, as well as the value of its data, change day by day and therefore, the controls need to be able to meet the new demands on a daily basis.

4.4.3 Data Governance Maturity Model

What is important to understand about Data Governance is that, it is a very complex and sensitive subject that cannot be implemented from one day to another. The implementation should be in small steps, so that the maturity level of the process and the people grows. There are several Data Governance Maturity level models. According to Nicola Askhan, Data Governance Coach, the Data Governance Maturity Model consists of the following steps (Fisher, T. (2009). The data asset: how smart companies govern their data for business success (Vol. 24). John Wiley & Sons.):

1. ***Unaware:*** In this initial level of maturity, there is no data management initiative, and the organisation does not understand the need to govern its data. The processes that take place during this phase, are unpredictable and poorly controlled, with no strict rules. Data might exist in multiple files and formats, stored across multiple systems with multiple names and no attempt to catalog has ever been made. Most of the companies nowadays, are beyond that level, since it is widely known the need of Data Governance.

2. ***Reactive:*** In this level, the scope is very limited and there are no specialist data governance or data quality tools. Instead, the organisation relies on a central person to implement the Data Governance. The success of the organisation at this level, depends on the technical analyst who is responsible for the “technical” aspects of the data.

3. ***Proactive:*** Moving on to the third level of maturity model, the whole culture of the organisation starts to change. Especially Financial organisations, who do not produce any other products than data, they really need to focus on them and treat them like assets. In the proactive stage, people have already started worrying about data, acquiring special tools for managing it, having a master data management (MDM) initiative. On this stage, the organisation is also looking for a single point of truth and starts investigating what is causing all the existing data issues. More importantly, people start understanding the importance of staff training, and IT and business start working together exchanging information about data. The challenge of this level, though, is that data and business processes remain separate, slowing innovation.

4. ***Managed:*** At the managed maturity level, all the data are being managed appropriately. The company does not need to manage all its data to the same level of management. It only has to manage it according to how critical it is. In order to classify your data according to their importance, you need to have risk controls and monitoring in place.

5. ***Optimized:*** This is the highest level any organisation can reach and in order to achieve that, all the previous steps need to occur step by step in order to help the business grow. Here, business is driven by standardized processes and manages to make the most value out of its data.

It depends on each company the extent to which they want to expand on the maturity level of Data Governance. Some organisations may not aim towards the optimized level of maturity model, for instance, but they can still use the model in order to move from the level they are currently in to the level they want to reach.

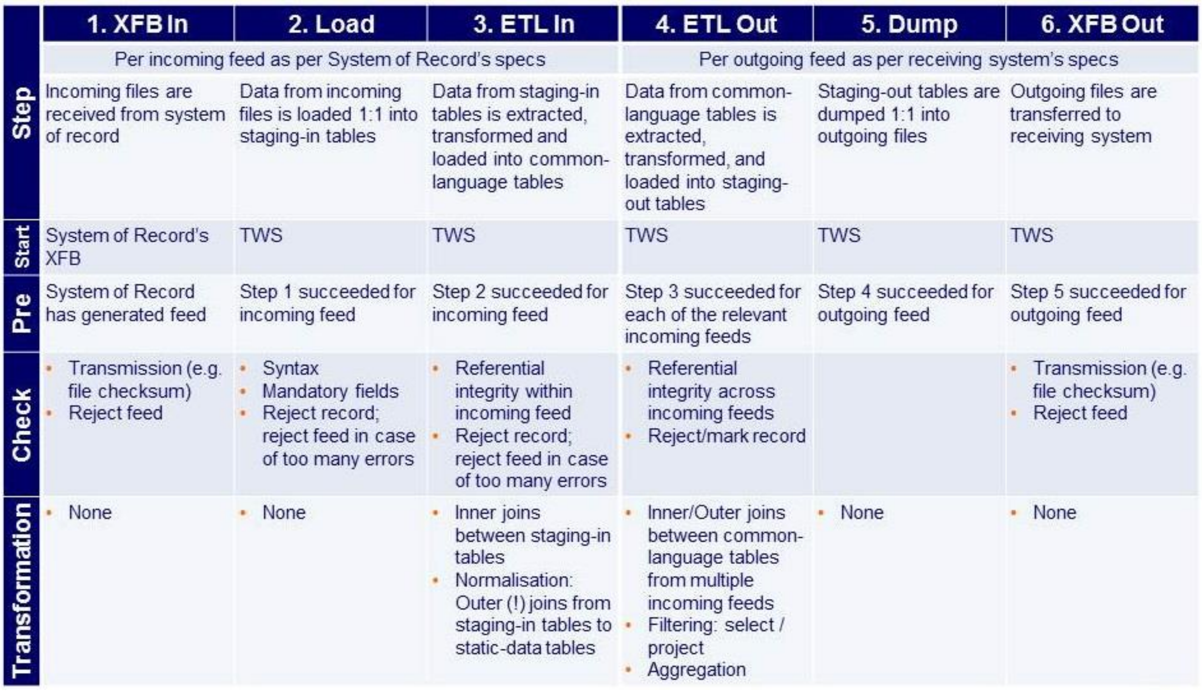
In this paper, the focus is mainly on the proactive and managed levels of maturity of Data Governance. Proactive is the stage to which the company currently belongs and the managed level is the target one. In other words, the Financial Institution started realizing that there is a swift in focus towards data and how to safely manage and protect them. It considers data as an asset and tries to train its employees towards MDM. What the organisation wants to achieve, is that managed maturity level of Data Governance, where the data is being appropriately handled and according to how critical it is by knowing its CIA ratings.

5 Research Context

5.1The Institution’s current situation

During the diagnostic phase of the research, the current situation of the organization is being examined.

The financial institution is currently releasing the Global Data Integration Layer (DIL), which is the first version of the data lake, and it is mandatory for every new post-transaction feeds from System of Records.. One of the first uses of the data lake, is the feed to Finance. DIL might also need to combine some of the data it receives from multiple resources into a single destination feed, creating Deep Data. In figure 4, the steps DIL has to follow are described in detail.



	1. XFB In	2. Load	3. ETL In	4. ETL Out	5. Dump	6. XFB Out
	Per incoming feed as per System of Record's specs			Per outgoing feed as per receiving system's specs		
Step	Incoming files are received from system of record	Data from incoming files is loaded 1:1 into staging-in tables	Data from staging-in tables is extracted, transformed and loaded into common-language tables	Data from common-language tables is extracted, transformed, and loaded into staging-out tables	Staging-out tables are dumped 1:1 into outgoing files	Outgoing files are transferred to receiving system
Start	System of Record's XFB	TWS	TWS	TWS	TWS	TWS
Pre	System of Record has generated feed	Step 1 succeeded for incoming feed	Step 2 succeeded for incoming feed	Step 3 succeeded for each of the relevant incoming feeds	Step 4 succeeded for outgoing feed	Step 5 succeeded for outgoing feed
Check	• Transmission (e.g. file checksum) • Reject feed	• Syntax • Mandatory fields • Reject record; reject feed in case of too many errors	• Referential integrity within incoming feed • Reject record; reject feed in case of too many errors	• Referential integrity across incoming feeds • Reject/mark record		• Transmission (e.g. file checksum) • Reject feed
Transformation	• None	• None	• Inner joins between staging-in tables • Normalisation: Outer (!) joins from staging-in tables to static-data tables	• Inner/Outer joins between common-language tables from multiple incoming feeds • Filtering: select / project • Aggregation	• None	• None

Figure 4: DIL detailed steps

It is already mentioned that the feed to Finance is one of the main uses of Data Lake. Figures 5 and 6, describe how the feed to Finance is being done right now within the company, and how the company plans to do it in the future, using the Data Lake.

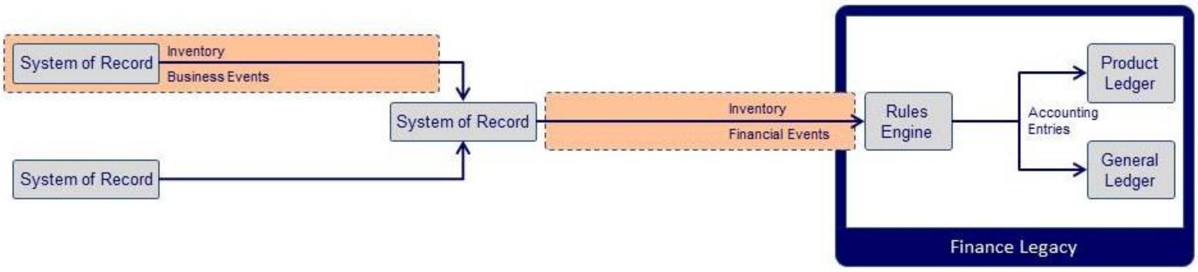


Figure 5: Current Feed to Finance

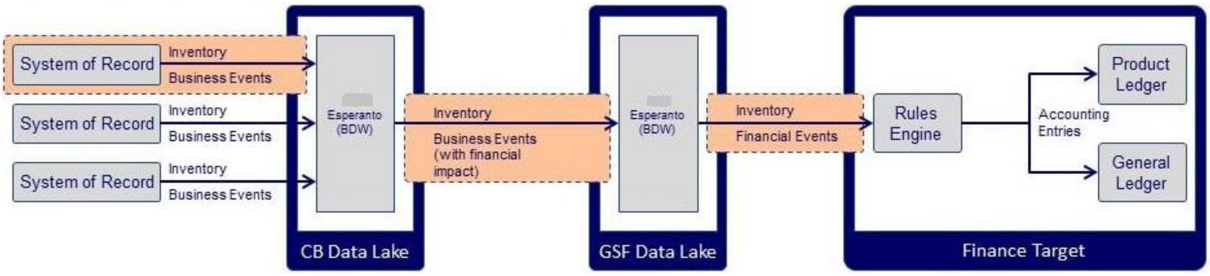


Figure 6: Target Feed to Finance

The Institution's plan about the implementation of the Data Lake is to build the template once but run many times and implement it properly per country and with standard installation guides. It is intended to contain an Extract-Transform-Load (ETL) framework that will help with the housekeeping and integration with the Catalog and the lineage. The catalog should describe all the data existing in the repositories inside the Data Lake together with its definition and origin (lineage).

5.2 Why does the Financial Institution need the Data Lake?

The Financial Institution understood the need of implementing a Data Lake solution, after it realized the great challenge of managing the great amount of data that it should deal with. The organisation adopted a Service Oriented Architecture (SOA) which made it successful in reuse of services across the whole company. The reuse of data assets however, became difficult as the organisation realized that data cannot always be discovered easily and in case it did, it was already replicated many times. The result of this process was both of low value of data and time consuming.

The interaction with customers is mainly based on the availability of great amounts of digitalized data which is being increased rapidly ad need to be of a high quality. The quality though, cannot be understood if the business department cannot access the lake easily. Many consumers depend on the same sources of data and therefore, having a common Data Lake will enable all the clients to use and combine the same data, while the distribution burden becomes lower as data only needs to be fed once into the lake.

What is more, the fact that all the data co-exist in the same area helps dealing with them on real time and making them more valid as the relevance of data depends on the context and timeliness. This also helps with the increasingly challenging security data assets, especially in case they need to be combined. Different data coming from different resourced and with different definitions but still, having to cover the needs of different lines of business, creates the need of combination of this data in order to be transformed towards a commonly comprehensive data model with standardized definitions.

5.3 IBM Launch Approach and Tooling for Data Governance in the Data Lake

The following paragraphs refer to the Action Planning phase of the Action Research method, where the actions and tooling needed in order to solve the existing problems of the Institution are being involved.

5.3.1 The Launch approach

In order to build the Data Lake, the Dutch Financial Institution decided to work together with IBM in order to successfully launch the Data Lake solution with the tooling IBM would provide it. IBM has a lot of years of experience in supporting business transformation projects, especially when it comes to analytics processes and is constantly improving its business operations in order to make better use of data and analytics.

According to IBM Corporation, there are certain steps the company needs to do before implementing the Data Lake Foundation described above. Those steps are to initially prepare the infrastructure, systems, software, etc., setup subscription tables, staging areas and code hub, design the usage of the catalog, identify the smallest possible environment the Lake runs on and prepare a demo approach from the user perspective. The demo scenario includes User stories for the next two phases, an explanation of IT Management and a data lineage from the business side. (Start the build of the Lake, May 13th, 2014,Smarter Analytics, IBM Corporation)

After following the steps described above, phase 1 begins. As illustrated in figure 7 below, the Data Lake operation steps are the following:

1. **Advertise** – After pre-filling the catalog with known business terms, give some extra role-based catalog information

2. **Discover** – Search the information in the catalog

3. **Catalog** - Describe all the data existing in the repositories

4. **Provision** data via information ingestion by modelling the repositories, defining mapping, storing raw data into Operational status and combined data into Information warehouse

5. **Explore** – Show the lineage and operational monitoring

6. **Access** – Publish information from both operational status and information warehouse repositories

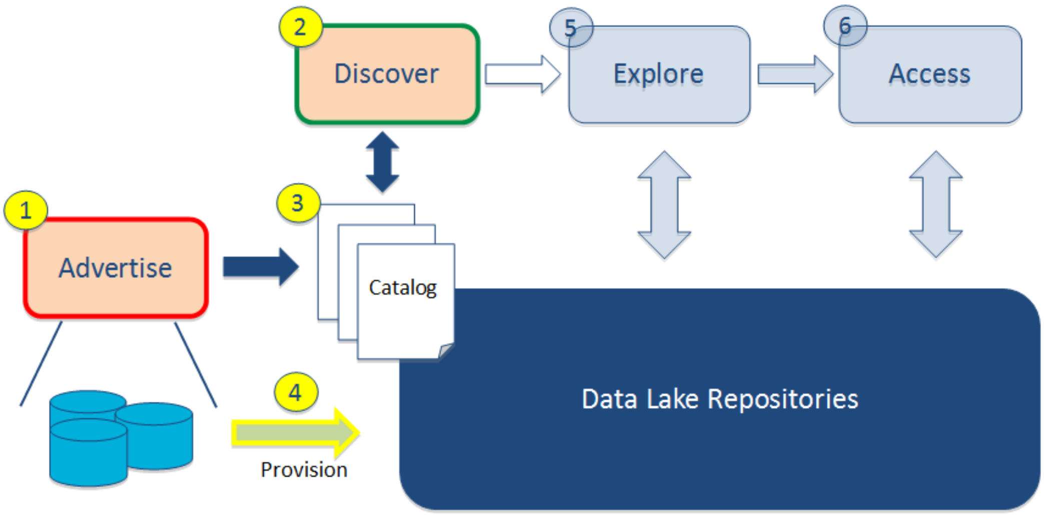


Figure 7: Data Lake foundation launch (Phase 1) – IBM Corporation, 2014

The second phase of the Data Lake foundation, is the implementation of the Big Data on top. This is done by loading data from Information Ingestion to the Deep Data repository and showing the lineage with the catalog and the Hadoop environment that is being used. (IBM Corporation)

5.3.2 The tooling

For successfully implementing Data Governance into the Data Lake, IBM provides the company with a tool called IBM Infosphere Server 11.3. This tool provides a single platform for information integration. It is used to manage metadata of all types (operational, business and technical) and helps people understand the relationships between the different meaning, structure and content of information across a wide range of sources through graphs and reports. More specifically, the tools that Infosphere Information Server provides the company in order to properly implement Data Governance into the Data Lake are the Infosphere Governance Catalog, the Infosphere Metadata Asset Manager, the Governance Dashboard as well as various consoles based on Data Stage and Quality Stage that increase capacity.

5.3.2.1 IBM Infosphere Governance Catalog

This tool is the most important part of the IBM Information Server when it comes to Data Governance. It provides an entry point for the company to understand and govern its information (metadata). Both IT and Business departments can create and manage information governance practices and a common business vocabulary. Within the catalog, users can create glossary assets: terms, categories, information governance policies, and information governance rules, and define relationships between the glossary assets that they create and other technical assets that they import. Besides the glossary, the Business Catalog also contains metadata, which are information about other internal or external data which are necessary for understanding the importance and sensitivity of the data the company has under its possession and manages. In addition, the catalog enables the company to explore and manage all these types of assets by providing reports on data flow, lineage, and the impact of changes to all these assets.

5.3.2.2 Infosphere DataStage

Infosphere DataStage is an ETL tool that helps the institution design data flows that extract data from a variety of sources, apply the transformation rules necessary in order to make it valuable, and load the data to the target application. (Alir, Takahashi, Toratani, & Vasconcelos, 2008) In order to develop this process, Data stage develops a graphical framework and, it uses parallel processing to provide a scalable platform.

The goals that the Institution could achieve by using Infosphere DataStage are described below:

- Design the data flows that extract information from multiple sources, transform this data using special transformation rules, and load the data to target databases
- Connect the applications of the Institution directly to the target systems to ensure that the data is relevant, complete, and accurate.
- Reduce development time and improve the consistency of design and deployment by using prebuilt functions.
- Minimize the project delivery cycle by working with a common set of tools across InfoSphere Information Server.

5.4 The company's awareness about Data Governance

In this phase, the evaluation of the research process is being presented. This includes a clear understanding of what people who work for DIL, as well as parties who come in contact with the Data Lake, think of the solution provided.

5.4.1 People's awareness

At the beginning of this research, several interviews were occurred, in order to get an insight of the employees' awareness of Data Governance and the reason for implementing it into the Data Lake. For this purpose, several people, from different applications used in the Data Lake have been interviewed, as well as the Data Lake Architects of different departments of the Organisation.

The questions were mostly focused on the understanding of people working for different Systems of Records or Receiving parties have about Data Governance. Those people were being asked about the nature of the data they use and send to the Data, Lake, their business value, their CIA (Confidentiality, Integrity, Availability) ratings, etc. They were also asked about the official roles people have in their department, such as an Information Owner, or somebody who is officially responsible for the data they handle, any other extra roles, and whether they have certain policies and standards, concerning the way they handle the data. Last but not least, those people were also asked about their department. How big their department is, how is the security awareness and education disseminated, the procedures they follow in order to evaluate their data, whether they use any particular program for backup and recovery, etc.

The Data Lake Architects were also asked some extra questions, useful for gaining a full insight of the company's awareness. Those questions are the following:

- Do you think that everybody in the DIL department is aware of their own roles and responsibilities?

- Do you have any specific policies and standards already in place?
- What is the program for backup and recovery of your data?
- What is the plan to keep different kinds of metadata up-to-date?

5.4.2 The Institution’s overall understanding

The organisation realizes how sensitive its data is and therefore, how crucial it is to protect all the information about customers, agreements, product usage, etc. in order to protect both its customers and market trust. The protection of the data is being achieved by the following procedures:

- Ownership should be allocated for all information Assets in the organisation
- For those information Assets, a CIA (Confidentiality, Integrity, Availability) rating should exist
- There should also certain Security policies and minimum standards be implemented. Corporate Risk Management should be responsible for those
- The minimum standards that are applicable should be applied per process in order to meet the required classes of CIA
- In case there are not available sufficient standards on technical level, or they are too expansive to implement, mitigations’ should compensate the risks that are not covered yet
- Also, in case those mitigations’ cannot close the remaining gap, risks need to be accepted by the accountable business owner
- Periodic reviews and audits lead to:

    1. *Re-evaluation of the ratings applicable*
    2. *Reporting of non-satisfactory covered needs*
    3. *Management of actions and change of planning to address the gaps that are open*

The company is already aware that the implementation and use of the Data Lake will not be easy as it lacks a lot of crucial factors that would help in the process. The main problems is that there is also no official information ownership and as a result, no clear responsibility about the data assets, no cost efficiency, no single source of truth. Furthermore, there is not yet a common data language to be used across all the departments that do transactions with the Data Lake. In general, since the organisation is at its very early stage of the implementation, there is a confusion about all the new concepts that people need to get used to. The most important part is to properly implement Data Governance into the Data Lake.

No access to the Data Lake should be given to any unauthorised party and there should be a Data Lake Security Mechanism to check the access rights. Even when a party is authorized to access the Data Lake, it still has to deliver the data in the agreed format in the Ingestion space. All these, need to be defined by the Data Lake itself. What is more, an Esperanto language should be defined across the departments as well as the other Data Lakes in the company in order to exchange information. Last but not least, a set of processes, roles and responsibilities are necessary to be implemented in order to manage data and data exchange between all parties and divisions that use systems, data lakes and reports, across the organisation.

6 Data Governance Implementation

6.1 Implementation on a higher level

As mentioned earlier, Data Governance is not an easy task. Bringing all the data together in one reservoir, may help the company enforce standards, but it could be hazardous as well. This is because, by keeping everything to one single place, enlarges the risk of losing everything as well. In case the security of the system gets compromised, then the company's only source of data is being misused and therefore, untrustworthy.

A second challenge that the company faces regarding the data Governance in the Data Lake is the adoption of the real-time processing. Such financial institutions, are mostly making use of batch-oriented solutions. The payments needed to be processed over a couple of days while banking, used to be done during office hours. In modern everyday routine though, people are using online payments, anytime, from anywhere in the world. As a consequence, the banking operation processes needed to change accordingly.

Another great challenge the organisation faces is the fact that the data are not easily accessible to the business users. In addition, I order for the institution to develop advanced analytics algorithms, it is necessary to give broad access to raw data, not only to business units, but to data scientists as well.

In order to overcome those challenges, and make the most of this Data Lake solution, it is essential that a proper Data Governance exists from the very beginning. The Data Lake is a completely new concept for the company and since it is on an early stage, a proper Data Governance needs to be implemented in order to start working properly and avoid confusions, duplication of data, etc. As we mentioned earlier, The maturity levels of Data governance are:

1. ***Unaware, 2. Reactive, 3. Proactive, 4. Managed and 5. Optimized*** and the company needs to move from the third to the fourth level.

More specifically, in DIL, there are specific rules and processes that need to be in place on a strategic, practical and operational level in order for the Data Integration Level to work properly.

6.1.1 Strategic laevel implementation

On a **strategic** level, the company has to investigate the existing situation in order to create a vision. Analysing the current and the target situation will help the organisation make a plan with the steps required in order to reach the ideal stage. This requires a proper evaluation of data and its quality with the proper tools, an estimation of the probability of risk and proper monitoring of the efficacy of controls. Figure 8 illustrates the strategic level of growth towards the target state.

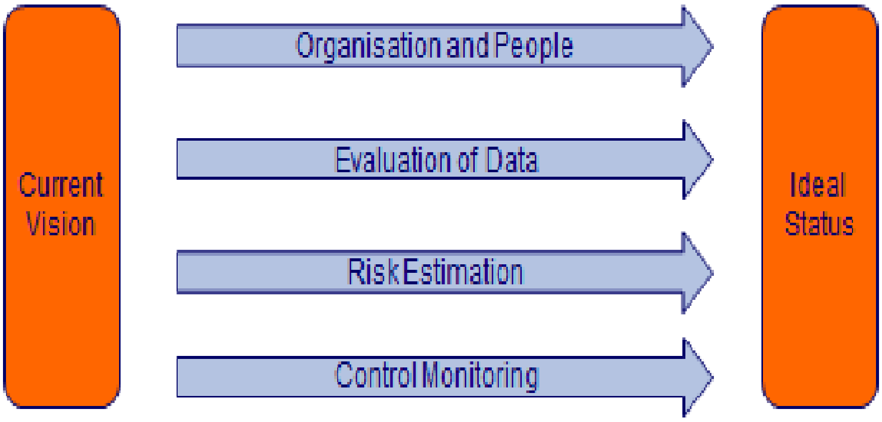


Figure 8: Strategic Plan

6.1.2 Tactical level implementation

On a **tactical** level, the organisation has to go through several steps in order to achieve the ideal state. When it comes to the Data Lake, the tactical steps are the following (o'Leary, 2014):

- ***Data Management and Warehouse:*** At this stage, the company has usually one source of structured data and it is developed to accommodate only a particular type of question which is already known upfront. It is a classic decision support method which Extracts, Transformsand Loads (ETL) data to an alternative database environment and the level of the data is on the level of the event occurring.

- ***Data Stage/Hadoop System:*** In this tactical level, we already have a Data Lake. A classic database approach in which users can access from multiple sources, combine raw data in order to make deep data, in both patch-based and real time.

- ***Real-Time Predictive Data:*** This last stage, is the ideal stage where the organisation would like to be. Especially nowadays, when the success of every business depends on how quickly it can react to conditions and trends, the ability of an organisation to analyse data in real-time is crucial.

In figure 9, an illustration of the tactical levels of analysis are being presented

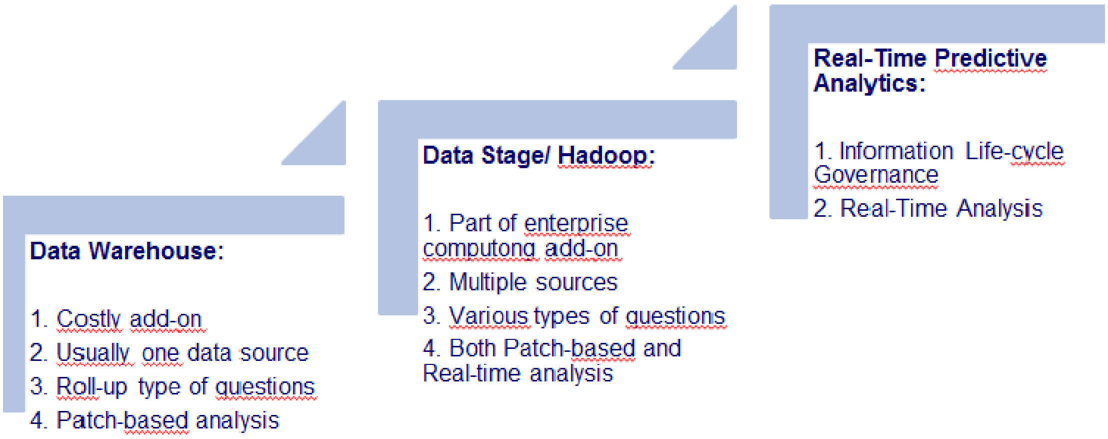
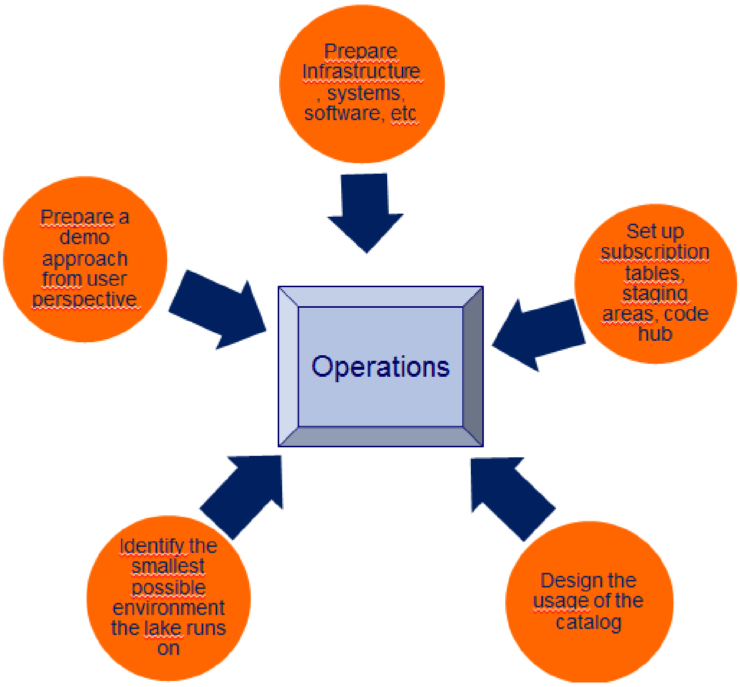

Data Warehouse:
1. Costly add-on
2. Usually one data source
3. Roll-up type of questions
4. Patch-based analysis
Data Stage/ Hadoop:
1. Part of enterprise computong add-on
2. Multiple sources
3. Various types of questions
4. Both Patch-based and Real-time analysis
Real-Time Predictive Analytics:
1. Information Life-cycle Governance
2. Real-Time Analysis

Figure 9: Illustration of tactical level

6.1.3 Operational level implementation

On an **operational** level, as mentioned already in 5.3.1, DIL needs to prepare the infrastructure, the systems and the software needed to support the Data Lake solution and to set up the subscription tables, the staging areas and the code hub. The next step should be to design the usage of the catalog and to identify the smallest possible environment the lake should run on. Last but not least, a demo approach from a user perspective should be prepared. This demo should contain user stories, the explanation of IT management and the data lineage in terms of business. (Start the build of the Lake, May 13th, 2014,Smarter Analytics, IBM Corporation)


Prepare Infrastructure , systems, software, etc
Prepare a demo approach from user perspective
Set up subscription tables, staging areas, code hub
Operations
Identify the smallest possible environment the lake runs on
Design the usage of the catalog

Fifure 10: Illustration of Operational Level

6.2Implementation on a lower level According to Scrum Agile Methodology

In order to ensure that Data Governance is implemented into the Data Lake successfully, the first thing the GDIL Department has to do is to create an Input and an Output model.

The Input model, consists of an Input Protocol that describes the Entry criteria according to which DIL will accept data into the Data Lake as well as the procedures the members of the DIL team should follow. The Output model, on the other side, consists of an Operational Level Agreement between the parties that come in contact with DIL in order to formalize the interface between the System of Records and the Receiving Application of the Solution.

6.2.1 Input Protocol

The Input Protocol, must be applied for all applications that are part of the Data Lake and want to either store or request data from the DIL department. The Entry Criteria that should be met by all applications are described below:

6.2.1.1 Entry Criteria

➢ ***Clear roles and responsibilities***

In order for DIL to accept data into the Data Lake, clear roles and responsibilities have to be strictly established. In case the source or the receiving application have not formally defined roles, DIL will not allow any further interaction as there should be a person responsible for the data that are being processed.

The *System of Records* needs to have a Data Owner and a Data Steward. The Data Owner is the person who is responsible and accountable for the data within the systems. He has all the legal rights and the complete control over the data which should be not used without his agreement. The Data Steward, on the other hand, is the person who assets the Data Owner in his everyday routine in order to ensure that data is properly managed and of the acceptable quality.

The application who receives the data, also known as *Line of Business Insight,* needs to have a Data User and a Data Custodian. The former, is the one who needs and requests the data from the Data Lake. In order to get the data he needs, he first needs to sign an agreement with the Data Owner who should be always on the System of Records side, in order to be able to get the data from the DIL. After he gets the data, he is allowed to further store, process or make reports out of them. The Data Custodian helps the Data User by providing the organisation and systems with data from one or more Data Owners.

➢ ***Clear File Descriptions***

❖ *File descriptions per source*

The filename:

- Short description of the file
- File type (binary os ascii)
- File pattern (ie testfile…) (start of the filename)
- Date format (yyyymmdd) (obligatory)
- Date format position (from which position in the filename)

- Time format (ie hhmmss); optional
- Time format position (from which position in the filename)
- Overview per recordtype (ie header, data, footer) including formats, length, primary key, decimal point etc.

❖ *File descriptions per target*

The filename:

- Short description of the file
- File type (binary or ascii)
- File pattern (ie testfile…) (start of the filename)
- Date format (yyyymmdd) (obligatory)
- Date format position (from which position in the filename)
- Time format (ie hhmmss); optional
- Time format position (from which position in the filename)

Overview per record type (ie header, data, footer) including formats, length, primary key, decimal point etc.

➢ ***Data is received in its raw format, without any modifications***

The Data received from the source systems should be delivered to DIL in its raw format, without any modifications. The purpose of the Data Lake is to store everything inside the lake as a single source of truth.

The receiving parties may need only a subset of the data and therefore, they can do any modifications needed (new business logic) in their side. The only change that should be possible in DIL is the combination of System of Records into the Data Stage.

➢ ***The BIA/CIA/P rating of the data is known***

The CIA ratings of the data received into the Data Lake should be known before the data enter the lake. The Data Lake could handle 3/3/3 data and, according to how sensitive the receiving data is, different actions should be followed.

➢ ***The quality of the data is being monitored***

Before any data enters the Data Lake, the DIL should be ensured that it is of a certain quality. The data should be accurate, complete and credible, as well as up-to-date.

➢ ***Compliance with the company's regulatory environment should be assured***

When it comes to sensitive data, it is crucial that any actions taken should be according to the rules of the legal and compliance department of t or regulators, as addressed in the OCD form.

6.2.1.2 Procedures inside DIL

In case an application complies with the Entry Criteria, the DIL department agrees to accept the data into the Data Lake. The model below describe all the detailed steps that should be followed from the moment there is a request for data to enter, until the data is finally stored into the Data Lake. All those procedures all also aligned with the scrum agile methodology used within the Financial Institution.

- ***Analyze the solution and data on demand***

In order for data to enter into the Data Lake, there is primarily an initiative either from the System of Records, or from the Receiving application (demand driven). The moment there is a request, the data on demand should be analyzed.

As mentioned above, the data need to be of certain quality in terms of accuracy, timeliness, completeness and credibility.

In addition, the data should be delivered in the DIL department in their raw format. Having access to the raw data available, provided the Financial Institution CB with the capability to have a single point of truth and develop advanced analytics algorithms.

- ***Identify Stakeholders***

The next step of the Intake solution, is to Identify the stakeholders have a role in managing the data exchange. There must be clear roles and responsibilities defined from each side of the information exchange before any further actions happen.

More precisely, a Data Owner should be assigned to be responsible for the data entering the lake. In case the request is demand driven, a data user from the receiving application should be assigned and responsible for the data and its transformation.

The role of Data Stewards is optional and at the discretion of the Data Owner whether to assign this role or not. Data Custodians are obliged to translate the requirements of the data users into the data requirements to the Data Owners as well as to arrange the delivery of data and the data quality agreements with both the Data Owners and the Data Users.

- ***Align solution with stakeholders***

After the roles from each side of the data transportation are clearly defined, the stakeholders should align their expectations of data use and authorizations. Every side should be clear about their expectations of the data use and transformation, as well as who is allowed and who is not to get access to the data. The Data Owner should have a formal agreement with the Data User, before the data is allowed into the Data Lake.

- ***Determine data rating***

The next step would be to identify the CIA ratings of the data entering the lake. According to those rating, the sensitivity of the data can be clarified and consequently, the necessary risk mitigation actions can be determined and realized.

- ***Agree on the data usage (IntakeLog)***

After following the steps above, the last step would be for the DIL department to take over the initiated solution as an epic into the “Intake Log”.

After having an “epic” in the product backlog, the Data Integration Layer (DIL) team needs to have a Backlog Refinement session. The purpose of this meeting is to estimate the effort required in order to successfully complete the epic, classify the requirements as well as divide the epic into smaller parts, called *user stories*. These procedures help the Product Owner of the team set priorities in the Backlog and plan tasks for the next Sprint.

The actions that have to be executed in the Refine Solution are the following:

- ***Determine the availability of the data:***

On this stage, both Information Owner and Data User have agreed on the conditions under which the data should be used and the DIL department starts the procedures required in order to obtain the data. The availability and accessibility of the data determines those actions. It usually depends on the place the data is stored, the people who are involved, the privacy policies, etc.

- ***Define the actions to obtain the data***

After determining the availability of the data, the whole team should discuss and define the actions needed in order to obtain the data and store it into the Data Lake.

- ***Create Data In, Transformation and Data Out Model***

After getting access to the data, DIL needs to check whether the data complies with the *Entry Criteria* mentioned in chapter 2. In case the data does not meet the necessary requirements, is should not enter the Data Lake. Thereafter, the DIL team should create a Data In, Transformation and Data Out model, according to the agreements between the System of Records, the receiving application, within the rules of CORM and Compliance of INV.

After the refinement meetings, the DIL team should follow some planning sessions in order for the Product Owner together with the whole scrum team (DIL) to agree on the Sprint goals and the priorities of the Product Backlog. The most important or urgent epics should be always on top of the Backlog board, refined for the next Sprint.

The steps that should be followed during the planning session are the ones below:

- ***Prioritize the solution in the product backlog***

The whole team should discuss and determine the priorities in the Backlog. It is important that the team works together and communicate with each other in order to stay focused on the tasks that have priority and not get distracted by other ones.

- ***Align with Stakeholders on timelines***

After setting priorities, the team should agree together with the stakeholders on timeliness in order to have an estimation on the time the product should be delivered.

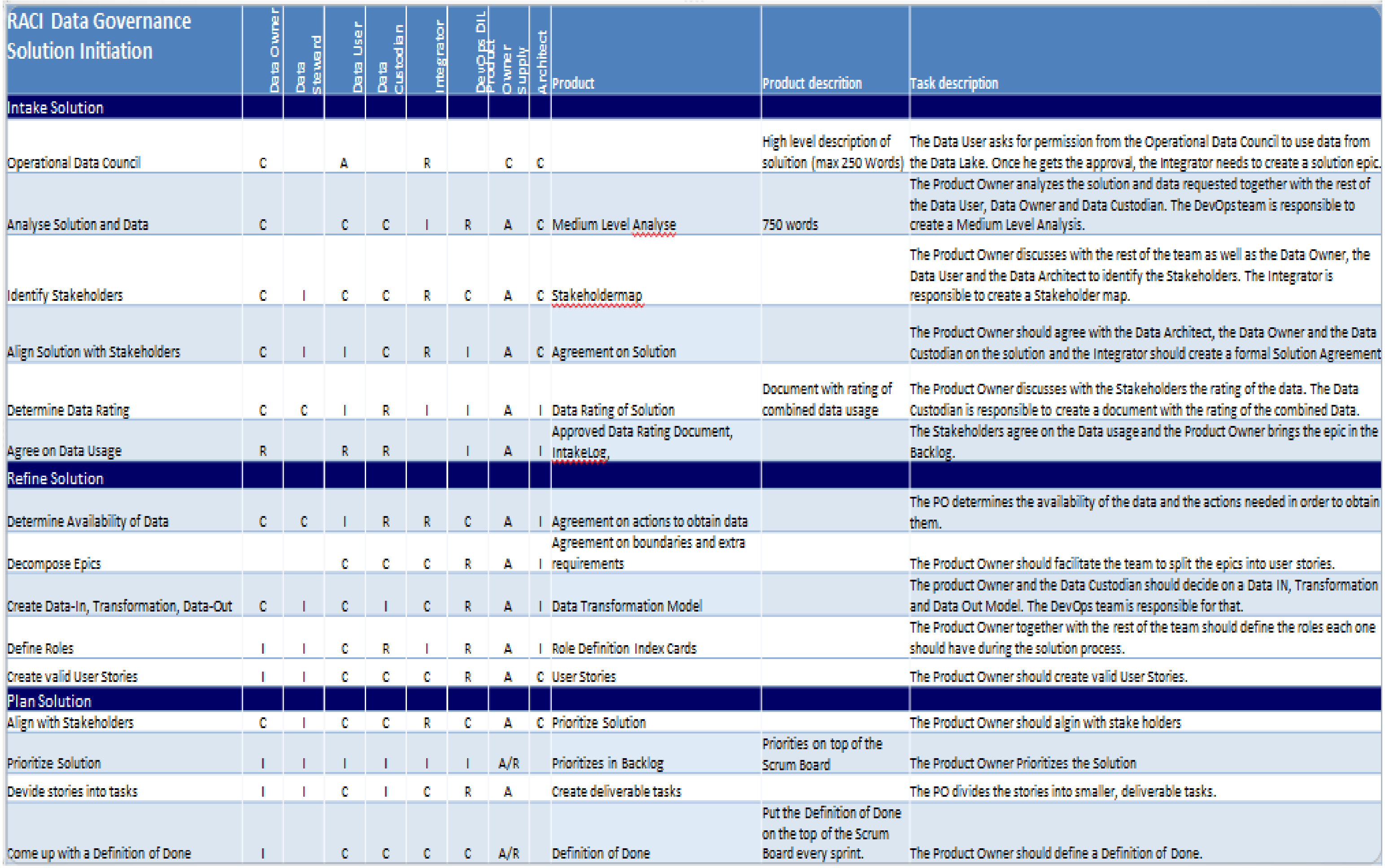
- ***Divide the stories into smaller tasks***

It is not necessarily expected from the team to know all the tasks that need to be completed from the beginning. Even though an estimation should be done, it is possible that during the sprint, other tasks such as unanticipated dependencies will be discovered.

- ***Come up with a Definition of Done***

What is also important after the planning session and before the sprint starts, is for the team to agree, together with the Product Owner on a Definition of Done. The reason such a definition should exist is to make clear and transparent for the whole team when is a product considered to be finished and ready to deliver. In order to avoid a “technical debt”, a product that is considered to be done, should be at least properly tested, refactored and potentially shippable.

The RACI chart below, indicates the roles of each member that takes part in the Initiation Solution:



RACI Data Governance Solution Initiation	Data Owner	Data Steward	Data User	Data Custodian	Integrator	DevOps DIL	Product Owner Supply	Architect	Product	Product descrition	Task description
Intake Solution											
Operational Data Council	C		A		R		C	C		High level description of soluition (max 250 Words)	The Data User asks for permission from the Operational Data Council to use data from the Data Lake. Once he gets the approval, the Integrator needs to create a solution epic.
Analyse Solution and Data	C		C	C	I	R	A	C	Medium Level Analyse	750 words	The Product Owner analyzes the solution and data requested together with the rest of the Data User, Data Owner and Data Custodian. The DevOps team is responsible to create a Medium Level Analysis.
Identify Stakeholders	C	I	C	C	R	C	A	C	Stakeholdermap		The Product Owner discusses with the rest of the team as well as the Data Owner, the Data User and the Data Architect to identify the Stakeholders. The Integrator is responsible to create a Stakeholder map.
Align Solution with Stakeholders	C	I	I	C	R	I	A	C	Agreement on Solution		The Product Owner should agree with the Data Architect, the Data Owner and the Data Custodian on the solution and the Integrator should create a formal Solution Agreement
Determine Data Rating	C	C	I	R	I	I	A	I	Data Rating of Solution	Document with rating of combined data usage	The Product Owner discusses with the Stakeholders the rating of the data. The Data Custodian is responsible to create a document with the rating of the combined Data.
Agree on Data Usage	R		R	R		I	A	I	Approved Data Rating Document, IntakeLog,		The Stakeholders agree on the Data usage and the Product Owner brings the epic in the Backlog.
Refine Solution											
Determine Availability of Data	C	C	I	R	R	C	A	I	Agreement on actions to obtain data		The PO determines the availability of the data and the actions needed in order to obtain them.
Decompose Epics			C	C	C	R	A	I	Agreement on boundaries and extra requirements		The Product Owner should facilitate the team to split the epics into user stories.
Create Data-In, Transformation, Data-Out	C	I	C	I	C	R	A	I	Data Transformation Model		The product Owner and the Data Custodian should decide on a Data IN, Transformation and Data Out Model. The DevOps team is responsible for that.
Define Roles	I	I	C	R	I	R	A	I	Role Definition Index Cards		The Product Owner together with the rest of the team should define the roles each one should have during the solution process.
Create valid User Stories	I	I	C	C	C	R	A	C	User Stories		The Product Owner should create valid User Stories.
Plan Solution											
Align with Stakeholders	C	I	C	C	R	C	A	C	Prioritize Solution		The Product Owner should algin with stake holders
Prioritize Solution	I	I	I	I	I	I	A/R		Prioritizes in Backlog	Priorities on top of the Scrum Board	The Product Owner Prioritizes the Solution
Devide stories into tasks	I	I	C	I	C	R	A		Create deliverable tasks		The PO divides the stories into smaller, deliverable tasks.
Come up with a Definition of Done	I		C	C	C	C	A/R		Definition of Done	Put the Definition of Done on the top of the Scrum Board every sprint.	The Product Owner should define a Definition of Done.

Table 2: RACI chart - Roles in the Initiation Solution

R= Responsible
A= Accountable
C= consulted
I= Informed

6.2.2 Output model

Besides the Services Procedures and Agreements, the Data Integration Layer (DIL) department should set a framework of the Operational Responsibilities that should be properly assigned and responsibly performed during the whole Solution process.

The Output model consists an Operational Level Agreement between the parties that come in contact with DIL. The aim of an OLA is, to formalize the interface between the System of Records and the Receiving Application of the Solution. The Data Integration Layer is only responsible for receiving and delivering data but holds no responsibility of the data itself. Therefore, the Data Lake cannot deliver any data before such an agreement is signed.

An OLA describes the interface the solution takes place in, the conditions of its operation, and the support that is being provided during the whole process. Moreover, it outlines the responsibilities of the parties being involved and aims to deliver a set of standard deliverables between them.

In order for an OLA to be complete, clear roles and responsibilities of all the parties involved should be assigned. Those Roles and Responsibilities are being described below:

➢ ***System of Records Responsibilities:***

The System of Records (SoR) is the application responsible for providing the Data Lake with data, once the Receiving Application requests it. The SoR System is responsible for the quality of the data, for delivering daily data feed to the Data Lake, to define the transformation to a common Esperanto language and, to archive the data. In more detail, SoR is responsible for the following actions:

- Guarantee the availability of the data within agreed time frames
- Communicate about the progress/status of incidents and problems
- Guarantee the continuity of automated processes of the interface
- Report any relevant changes in the production environment
- In case a file is delivered with the wrong data because of an issue in a system, SoR is informed and should resend the file with the correct format and data.
- Every EOD (End Of Days) delivery will be delivered separately. It will not be clustered into one delivery in case of issues.
- Changes in the data will be limited as much as possible as the data need to be placed in the Data Lake in its raw format

➢ ***Receiving Application Responsibilities:***

The receiving application is usually the one initiating the whole process (demand driven). In General, the Receiving System is responsible for data consolidation, receiving (daily) data feed from the Data Lake, defining the transformation from the common Esperanto Language and building historic data. In particular, in an operational level, the Receiving System is responsible for the following actions:

- Assisting in guaranteeing the continuity of automated processes required to obtain the data from the System of Records
- Guaranteeing the availability of the interface

- Reporting any possible incidents or problems
- Reporting any possible changes in their production environment
- Reporting any possible changes in the usage of the data they are receiving
- Investigating and planning to implement notifications to the System of Records in case they do not receive the requested data within the timelines agreed

➢ ***DIL Responsibilities:***

The main role of the Data Lake is to receive data from the System of Records, deliver it to the Receiving System and, implementing the transformations to/from the common Esperanto language. In an operational level, DIL should follow the steps below:

- Guarantee the availability of the data within agreed time frames
- Communicate about the progress/status of incidents and problems
- Guarantee the continuity of automated processes of the interface
- Communicate about relevant changes in the production environment
- In case that a file is delivered with the wrong data because of an issue in a system, the SoR is informed and should resend the file with the correct format and data.
- Every EOD delivery will be delivered separately. It will not be clustered into one delivery in case of issues
- Changes in the DIL data will be limited as much as possible and will be stored in a different place within the lake, where all the deep data is being stored

The RACI chart below indicates who is accountable or responsible for every action:

	Data Owner	Data Steward	Data User	Data Custodian	Integrator	DevOps Team	Product Owner	Architect
Guarantee availability of Data	C		C		R		A	
Report status of incidents	I		I			R	A	
Guarantee continuity	I		I			R	A	C
Report changes in production	I		I	C	R		A	
React in	I		C		R	R	A	

case of wrong data								
Separate EOD deliveries						R	A	
Changes in data	I		I	C	A	R	A	

Table 4: RACI char - Action Roles

R= Responsible
A= Accountable
C= consulted
I= Informed

7. Conclusion – Further Research

This thesis intended to introduce a conceptual framework for implementing Data Governance into the Data Lake in the context of a Big Dutch Financial Institution. The Data Covernance implementation was explained in both higher level and, most importantly, lower level, aligned with the Scrum methodology. Hence this research provided a useful conceptual framework that is not only already being followed by the Financial Institution, but could be also used by other organizations as well.

The use of the conceptual Framework, has helped the DIL team follow certain steps that were not clear before in order to properly implement Data Governance into the Data Lake. This is really important, especially in this very early phase, because the earlier they start implementing Data Governance properly, the easier it will get later on. In addition, the people in the team, especially the DevOps team is no longer confused and understands the reason behind every action they take. Having the whole picture in their mind, and who is responsible for each part of the procedure, helps effectiveness and collaboration in the team and between teams.

Despite the fact that the results look promising enough, and that the framework is already being used by the Institution, research on Data Governance should never stop. Data are becoming more and more important for organizations and since we live in a world that is continuously being changed, different frameworks might be needed later on, more suitable for the future circumstances.

References:

Agile/Lean Data Governance Best Practices (Agile/Lean Data Governance Best Practices) http://agiledata.org/essays/dataGovernance.html#Traditional

Alir, N., Takahashi, C., Toratani, S., & Vasconcelos, D. (7). IBM Infosphere DataStage Data Flow and Job Design (p. 658). ISBN-13: 9780738431116, ISBN-10: 0738431117

Anderson, J., Aydin, C., & Jay, C. (1994). Qualitative Research Methods for Evaluating Computer Information Systems. In Evaluating Health Care Information Systems: Methods and Applications (pp. 45-68). Sage, Thousand Oaks, CA.

Asquer, D. A. (2013, January, 1). 34. The Governance of Big Data: Perspectives and Issues. ICCP 2013 Conference: First International Conference on Public Policy. Financial and Management Studies, SOAS, University of London

Avison, D., Lau, F., Myers, M., & Nielsen, P. (1999). Action research. Communications of the ACM, 42(1), 94-97

Baskerville, R. (1993). Information systems security design methods: Implications for information systems development (4th ed., Vol. 25, pp. 375-414). ACM Computing Surveys (CSUR).

Big data and positive social change in the developing world: A white paper for practitioners and researchers. (2014). Bellagio Big Data Workshop Participants, Oxford: Oxford Internet Institute.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. Provocations for a Cultural, Technological and Scholarly Phenomenon. Information, Communication & Society, 662-279.

Cate, F. (2010). Data Tagging for New Information Governance Models. IEEE COMPUTER AND RELIABILITY SOCIETIES. doi:1540-7993/10/$26.00 © 2010 IEEE

Chessell, M., Nguyen, N., Van Kessel, R., & Van Der Starre, R. (2014). Governing and Managing Big Data for Analytics and Decision Makers. IBM Redbooks.

Crisan, C., ZBUCHEA, A., & MORARU, S. (2014). Big Data: The Beauty or the Beast. Management, Finance, and Ethics. doi:10.13140/2.1.2709.7282 Conference: Strategica 2014

David, T. (2006). American Journal Of Evaluation. 27(2), 237-246.

Dhillon, G., & Backhouse, J. (2001). Current directions in IS security research: Towards socio-organizational perspectives. Information Systems Journal, 11(2), 127-153.

Fisher, T. (2009). The data asset: How smart companies govern their data for business success (Vol. 24). John Wiley & Sons.

Fu, X., Wojak, A., Neagu, D., Ridley, M., & Travis, K. (2011). Data Governance in predictive toxicology: A review. Journal of Cheminformatics.

James, M. (2011, April 15). The Backlog Refinement Meeting (or Backlog Grooming).

Khatri, V., & Brown, C. (2010). Designing Data Governance. Communications of the Acm, 53(1).

Leedy, P., & Ormrod, J. (2005). Practical research. Upper Saddle River, NJ: Prentice Hall.

Legal aspects of managing Big Data. (2015). Cpomputer Law & Security Review, 31(1), 1-169.

Levin, D. (n.d.). The opening of vision: Nihilism and the postmodern situation.

Lewis, I. (1985). Social Anthropology in Perspective.

Loshin, D. (2013). Data Governance for Master Data Management and Beyond. SAS Institute Inc. World Headquarters White Paper.

Martin, P., & Tumer, B. (1986). Grounded Theory and Organizational Research. The Journal of Applied Behavioral Science, 22(2), 141-157.

May, T. (1997). Social research: Issues, methods and process (2nd ed.). Trowbridge: Redwood Books.

Mayers, M. (2013, September 3). Qualitative Research in Information Systems. Association for Information Systems (AISWorld) Section on Qualitative Research in Information Systems, 241-242.

Otto, B. (2011). A MORPHOLOGY OF THE ORGANISATION OF DATA GOVERNANCE. ECIS 2011 Proceedings.

O'Leary, D. (2014). Embedding AI and Crowdsourcing in the Big Data Lake. University of Southern California.

Rapoport, R. (1970). Three Dilemmas in Action Research. In Human Relations (Vol. 23:6, pp. 499-513).

Roland, R. (1985). Research Methods in Information Systems (p. 193201). Amsterdam, NorthHolland.

Russom, P. (2006). Taking data quality to the enterprise through data governance. The Data Warehousing Institute.

Sarsfield, S. (Director) (2015, February 2). Data Governance Imperative. Cambs, GBR: IT Governance. Lecture conducted from ProQuest ebrary, .
Schiffman, L., & Kanuk, L. (1997). Consumer Behaviour. London: Prentice Hall.

Seiner, R. (2012, December 1). 40. Applying an Maturity Model to Data Goernance. The Data Administration Newsletter.

Simonsen, J. (2009, January 1). [Radio broadcast]. Scandinavia: Molde University College. IRIS 32, Inclusive Design. (pp. 1-11).

Siponen, M. (2002). Designing secure information systems and software: Critical evaluation of the existing approaches and a new paradigm.

Sprint Planning Meeting. (2014, August 20).

Start the build of the Lake. (2014, May 13). Smarter Analytics, IBM Corporation.

Tallon, P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. Loyola University Maryland, 13, 0018-9162.

Taylor, P., Nevitt, C., & Carnie, K. (2012, December 11). The rise of Big Data. Financial Times.

The IBM Data Governance blueprint, Leveraging best practices and proven technologies. (2007, May 1).

Trope, R., Power, E., Polley, V., & Morley, B. (2007). A Coherent Strategy for Data Security through Data Governance. 1540-7993.

Weber, K., Otto, B., & Osterle, H. (2009). One Size Does Not Fit All – A Contingency Approach to Data Governance. ACM Journal of Data and Information Quality, 1(1).

Wende, K. (2007, December 5). [Radio broadcast]. Tonwoomba: Kristin Wende.

Wood, C. (1990). Principles of secure information systems design. Computers & Security, 9(1), 13-24.

Yin, R. (2002). Case Study Research, Design and Methods

Zhu, J. (2011). Metadata Management with IBM InfoSphere Information Server. IBM Redbooks.

O'Leary, D. (2014). Embedding AI and crowdsourcing in the Big Data Lake. AI Innovation in Industry, University of Southern California.